DOCUMENT RESUME

ED 110 493 TM 004 766

AUTHOR Strassberg-Rosenberg, Barbara; Donlon, Thomas F.
TITLE Content Influences on Sex Differences in Performance on Aptitude Tests.
PUB DATE [Apr 75]
NOTE 45p.; Paper presented at the Annual Meeting of the National Council on Measurement in Education (Washington, D.C., March 31-April 2, 1975)

EDRS PRICE MF-$0.76 HC-$1.95 PLUS POSTAGE
DESCRIPTORS Academic Aptitude; *College Admission; Comparative Analysis; *Item Analysis; Mathematics; Senior High Schools; Sex (Characteristics); Sex Differences; *Sex Discrimination; *Standardized Tests; *Test Bias; Testing Problems; Tests; Verbal Tests
IDENTIFIERS *Scholastic Aptitude Test

ABSTRACT

The purpose of the present study is to examine the April, 1974 Scholastic Aptitude Test (SAT) for item-content bias between the sexes. By so doing, this study forms a logical extention of the work of Coffman (1961) on the '54 SAT, and Donlon (1973) on the '64 SAT. A study of item-sex bias was conducted using the method of delta-plots (Angoff, 1972; Angoff & Stern, 1972). Those items demonstrated to have different "psychological meaning" were then investigated for patterns of content bias by referencing to the test assembler's classifications. In addition, the test was inspected using the criteria established by Tittle, et. al. (1974) and Lockheed-Katz (1973) for determining sex bias. The results of the two methods of analysis were compared. (Author)

****
* Documents acquired by ERIC include many informal unpublished *
* materials not available from other sources. ERIC makes every effort *
* to obtain the best copy available. nevertheless, items of marginal *
* reproducibility are often encountered and this affects the quality *
* of the microfiche and hardcopy reproductions ERIC makes available *
* via the ERIC Document Reproduction Service (EDRS). EDRS is not *
* responsible for the quality of the original document. Reproductions *
* supplied by EDRS are the best that can be made from the original. *
****

ED110493

Content Influences on Sex Differences in Performance on Aptitude Tests

Barbara Strassberg-Rosenberg
State University of New York at Buffalo

Thomas F. Donlon
Educational Testing Service

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION
THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY

TM004766

NCME
1975

Abstract

The purpose of the present study is to examine the April, 1974 Scholastic Aptitude Test (SAT) for item-content bias between the sexes. By so doing, this study forms a logical extention of the work of Coffman (1961) on the '54 SAT, and Donlon (1973) on the '64 SAT. A study of item-sex bias was conducted using the method of delta-plots (Angoff, 1972; Angoff & Stern, 1972). Those items demonstrated to have different "psychological meaning" were then investigated for patterns of content bias by referencing to the test assembler's classifications. In addition, the test was inspected using the criteria established by Tittle, et. al. (1974) and Lockheed-Katz (1973) for determining sex bias. The results of the two methods of analysis were compared.

Content Influences on Sex Differences in Performance on Aptitude Tests

Barbara Strassberg-Rosenberg[1]
State University of New York at Buffalo

Thomas F. Donlon
Educational Testing Service

Maccoby and Jacklin's (1974) survey of the literature suggested that women are more verbal, men more quantitative -- but is this true, or is it possibly an artifact of item-sex bias in the standardized examinations which were studied, and cultural tracking of the sexes into socially acceptable stereotypical courses? At least one major test, the College Entrance Examination Board's SAT-Verbal, does not reveal this "established" pattern, for on this test the sexes score equally well. Is this indicative of a change in the comparable intellectual ability levels of the sexes, or possibly, a different balance of sex-interactive content within the test? The primary objective of the present study is to investigate the possibility of sex bias in the items of the April, 1974 form of the Scholastic Aptitude Test (SAT). Item-sex bias is statistically defined in this paper as an aberration in the pattern of difficulty for males or females on a specific item when that item is compared to performance on other items within the test or section. That is, those individual items which demonstrate an item-group interaction effect, as determined by an analysis of the plots of item difficulty,

[1]The authors would like to acknowledge the contributions of Dr. T. Anne Cleary, of CEEB, who supported the study; Ms. June Stern of ETS, who consulted on a number of statistical matters; and Dr. Jeremy D. Finn, of SUNY-Buffalo, who read and criticized an early draft.

are considered as sex-biased. Thus, content bias is considered, if an apparent content factor can be determined for the biased item. Language bias, as well as stereotyping (Tittle, McCarthy, & Steckler, 1974; Lockheed-Katz, 1973) is also taken into account. Basically, the authors have attempted to analyze the items in both a subjective and an objective sense, exploring in depth the potential sources of bias in the individual item.

This inquiry forms a logical and timely extension of the work of Coffman (1961) on sex differences in performance on the 1954 SAT and Donlon (1973) on content factors in sex differences on the 1964 SAT. Coffman compared the performance of a male and a female sample (male=370, female=370) on the verbal aptitude section of the March, 1954 Scholastic Aptitude Test. Drawing off the top and bottom 100 cases from each sample, item difficulties were computed as values of ϕ using an arcsin transformation suggested by Walker and Lev (p. 423-24). Differences between corresponding ϕ's for the male and female samples were obtained. While the total verbal score indicated no differences in general verbal ability between the male and female samples, individual items were less uniform. Those showing large differences in difficulty were inspected for possible content explanations. Thus Coffman categorized the three items on which the women's sample did better as involving words which describe personal feeling or personality characteristics. Of the six items favoring the male sample, on the other hand, three concerned mechanics or business vocabulary. Coffman could not develop a content

hypothesis for three other items. In another phase of the study, when a test construction specialist studied the 60 verbal omnibus[2] items, independent of the item analysis, he discovered 17 items for which he predicted a significant sex difference in performance and the direction of this difference. In 14 cases he was correct. Coffman concluded that if "....differences which appear in responses to aptitude test items are relevant, then items showing differences should be removed from the item pool or controlled at the point of assembly to insure optimum weighting in test" (Coffman, p. 124).

Donlon studied the performance of the entire population of candidates for the May, 1964 Scholastic Aptitude Test. For the SAT-Verbal, this involved the scores of 55,717 males and 47,083 females; while for the SAT-Math, the scores for 55,717 males and 47,082 females were available. For total test scores, no performance difference was found between males and females on the verbal. Males, however, were superior on the math. Donlon followed Coffman's method of investigating items demonstrating large differences between item difficulties, using as the index of item difficulty "p" - the proportion of candidates reaching the item who answered it correctly. For the verbal section, Donlon found that items classified by the test assembler as Human Relationships, Humanities or Aesthetic - Philosophical are easier for females, while items classed as World of Practical Affairs or Science are easier for males. This corroborates Coffman's findings that items involving words related to "people" are easier for women, while items relating to "things"

[2]Verbal omnibus items include both the discrete verbal and reading comprehension items, thus composing the total verbal section.

are relatively easier for men. Donlon also included summaries of the average differences in item difficulty (p-differences) by item type. He found that "....only the sentence completion material was truly balanced in this form. Antonyms and analogies tended to favor females and males respectively, by equal amounts, while reading comprehension favored females by the largest amount" (Donlon, p. 10). In his analysis of the mathematical section, Donlon found that, on the average, each math item favored males by about .07. Only two items demonstrated a difference in favor of females. Of these two items, one involved a household setting while the other was an "algebra" question. Surveying the 60 math items, he found that 17 has real world referents (e.g., pulleys, wheels, cars, etc.), while 19 items could be classified as "algebra." The report comments that "There seems to be a masculine tenor to the contents of the 17 "subject matter" items. No females are agents in this world. We meet 'a boy,' 'John,' 'a man,' 'Mr. Brown'.....Nor are these feminine things" (p. 14). Comparing male-female performance on the "subject-matter" versus "algebra" questions Donlon states, "If only 'subject matter items' were used, the male 'advantage' could grow to about 60 scale units. If the items were limited to 'algebra' the differences could diminish to about 20 points" (p. 16). While the study cannot establish that content is the determining factor in the difference in item performance between males and females, Donlon concludes with the advice that "Long-standing and stereotyped expectations of sub-group performance may be less permanent that is believed" (p. 18).

Data Source

The Scholastic Aptitude Test scores are based upon two separately timed math sections. While additional sections are included for equating and pretesting, only those score-producing sections were considered in the present analysis. The verbal section consists of discrete verbal and reading comprehension questions. The "discrete" items are so called because "they are complete in themselves, rather than being associated in common with a passage as are reading comprehension questions" (Donlon & Angoff, 1971, p. 22). The discrete verbal items consist of three item types: sentence completion, antonyms, or analogies. Each item type is also coded by content as Aesthetic-Philosophical, World of Practical Affairs, Science, Human Relationships, or General. For the reading comprehension items the content is determined by the content of the passage. "There are currently seven passages, each with five associated questions, in a typical SAT. These passages consist of one each from the following seven content categories: narrative, biological science, physical science, synthesis, argumentative, humanities, and social studies" (Donlon & Angoff, 1971, p. 22). The math items are formally divided into two major item types: regular math and data sufficiency. In a data sufficiency question, the candidate need not solve the actual problem, but must simply decide if sufficient information has been given to solve it. The math items are further classified by content as Algebra, Geometry, Arithmetic, or Miscellaneous.

Method

From the population of 449, 266 candidates who took the Scholastic Aptitude Test in April, 1974, a random sample of 5,993 examinees was available. Of these, one thousand subjects of each sex were selected by a random-spaced sampling technique for the present study.

An item analysis was performed for the male and female groups separately. The proportion of each group responding correctly to each item (the p-value) was computed and transformed to a normal deviate, delta[3]. The delta-values for the female group were then cross-plotted with the delta-values for the male group, resulting in an elliptical pattern of points for the set of items which was compared. "....the correlation coefficient represented by the ellipse represents the degree to which the items have the same rank order of difficulty in the two groups -- also a representation (inversely) of the item x group interaction" (Angoff & Stern, 1971, p. 7). Such delta-plots were constructed for the verbal omnibus (total verbal), total math, ~~discrete verbal (vocabulary including sentence completion, antonyms, and analogies), reading comprehension~~, regular math, and data sufficiency items on the test. The major axis of the ellipse for each delta-plot was algebraically determined by a procedure developed by Angoff & Stern (1971, p. 7-8) and the perpendicular distance (D) of each item-point from it computed. The standard deviation of the distribution of these distances

[3]Delta (Δ) is an index of item difficulty for an item. $\Delta = 4z + 13$, where z is a normal deviate corresponding to p, the proportion of examinees answering the item correctly (Angoff & Stern, 1971).

is a function of the item x group interaction (Angoff & Ford, 1971, p. 5; Angoff & Stern, 1971, p. 7). The items departing most extremely from the concentration of points in the plots (outliers) are regarded as contributing to the item x group interaction. They are the items that do not fall in the same rank order of difficulty for the males and females. Such items seem to represent a different "psychological meaning" (Angoff, 1972) to the members of the two sexes, and are defined as sex-biased in this study. "They are the items that are especially more difficult for one group than the other, relative to the other items....." (Angoff, 1972, p. 2). That is, when the distance from the major axis is less than zero, the items tend to be more difficult for the group defined on the ordinate than were most other items of the same test for the same group. Whereas, if the distance is greater than zero, the item tends to be more difficult for the group defined on the abscissa than were most other items on the same test for the same group. If the distance equals zero, the item is of approximately equal difficulty for the appropriate group as compared to the other items on the test.

The test assembler's formal classifications of items were examined to see if a relationship exists between item content and different "psychological meaning." The individual items of the tests were also inspected using the criteria developed by Tittle, McCarthy, and Steckler (1974) and Lockheed-Katz (1973) in regard to language bias and sex stereotyping. Ratios of the frequency of usage of male nouns and pronouns to female nouns and pronouns were

computed for each section. The status of the male and female characters within the test were compared as a check for equality. The items were further analyzed to determine if their content reinforced traditional or stereotyped images of men and women.

Results and Discussion

The means and standard deviations of the raw scores and of the item difficulty indices can be found in Table 1 for both the male and female samples on the verbal and mathematical sections of the

---

Insert Table 1 about here

SAT. The verbal scores are virtually equivalent for the male and female samples. That is, neither the males nor females, on the average, appear to have greater verbal ability as measured by performance on the SAT. The differences between means do not exceed that which would be expected by chance ($p.>05$) for both the item and score data on the verbal section. This finding is not consistent with the Maccoby and Jacklin conclusion that girls and women are higher in verbal ability (1974). The same finding surfaced in the Coffman and Donlon studies, apparently indicating a trend towards equivalence in male-female performance on the SAT-Verbal. This equivalence was not discernible in the earliest forms of the test (Donlon, 1974). As can be seen in Table 1, the males did score higher on the mathematical sections ($p<.01$). Both the regular math

and data sufficiency subtests proved to be more difficult for the females than for the males as illustrated by the higher female Δ's ($p<.01$). This confirms Donlon's finding of superior male mathematical performance on the SAT and substantiates Maccoby and Jacklin's conclusion that males are more mathematically able.

The intercorrelations between SAT total test scores and subscores for the male and female samples appears in Table 2. Of

---

Insert Table 2 about here

---

interest is the fact that a higher correlation exists between the regular math and data sufficiency subtest for the males than for the females ($p<.01$). That is, performance on the regular math and data sufficiency sections seems to be more related for males than for females. This is of interest in light of Donlon's finding of a slightly improved position for females in respect to males on the data sufficiency items. Referring to Table 1, this data also suggests that the males do not have quite as great an advantage over the females on the data sufficiency scores as in the regular math section.

As discussed above, males and females do about equally well on the SAT-Verbal, and males perform better on the SAT-Math. Overall performance, however, does not answer the question of whether item content bias is present. A group may perform better or worse on an aptitude measuring instrument because of different abilities; but

it is also possible that overall performance can be unknowingly weighted to favor a particular group by the specific content of the individual items. In effect, total scores might be masking important item performance variance caused by bias, rather than ability. Thus, the focus of the following analysis will be on the individual item, rather than on the total score.

Table 3 gives the distribution of observed D-values -- the orthogonal distances from the points on the ellipse to the major

---

Insert Table 3 about here

---

axis of the ellipse -- for the item delta plots for verbal, total mathematical and separately for two classes of mathematical material. In addition, a statistical summary of the results is presented. The correlations between deltas for the male and female samples are all greater than .9, indicating a high correspondence between the rank orders of item difficulties on the various sections for these groups. The standard deviations of the D-values are correspondingly small. It should be noticed that the D-values have a mean of zero (0). Negative D-values are measures to points that lie <u>above</u> the major axis line. As can be seen in the delta plots, Figures 1-4, the male delta values are plotted along the abscissa (X axis) and the female

---

Insert Figures 1-4 about here

---

delta values are plotted on the ordinate (Y axis). A dotted line

has been drawn at a 45° angle to serve as a reference. Items falling on or near this reference line are of approximately equal difficulty for the male and female samples. Negative D-values, distances to points above the major axis, are easier for the male group, but more difficult for the female sample relative to other items in the particular test or subtest. Conversely, positive D-values tend to be easier for the female group and more difficult for the male group relative to the rest of the test. An inspection of the frequency distributions (Table 3) reveals that the largest negative values are greater than the largest positive values for all sections, indicating a balance of specific items biased in favor of males or biased against females. The greater the distance from the major axis (the greater the absolute D-value) the more the particular item contributes to an item x group interaction. For purposes of this study, it was decided that all D-values beyond ± 1.5 standard deviations would be investigated for possible sex-content bias. For example, for the total verbal (verbal omnibus) the standard deviation of D = .3959, and all items with a absolute D-value greater than 1.5 x .3959 or .5939 were investigated.[4]

Insert Table 4 about here

Table 4 lists the item number, D-value, and normalized D-value (D-value divided by the standard deviation of the D's) for all verbal items. For the total verbal (N = 90), 12 items met the criterion.

---

[4]This can also be found by taking all items with a normalized D-value ± 1.5 (Table 4).

(starred items, Table 4). These items are also labeled in Figure 1, the delta plot for the SAT verbal. As can be seen, nine items were biased towards males (above line, negative D-value), while three were biased in favor of females (below line, positive D-value). These twelve items are listed in Table 5 with their D-values, normalized

---

Insert Table 5 about here

D-values, difficulty indexes, assemblers code, and item type. Of the nine items biased towards males, four were coded by the test assemblers as World of Practical Affairs, and three were coded as Science. The remaining two were reading comprehension items, one from a Science passage, the other in a Synthesis passage with historical-political tones. As the difficulty index (Δ) of item 39 (R.C. - Synthesis) is so high for both males and females, and as 39 is the next to the last item in Section I, it might be argued that candidates were guessing in haste based on partial information. Thus, there is the possibility that this item is not really biased - but just rushed on. Of the three items in favor of women, two were classified as Aesthetic-philosophical, while the third was coded as World of Practical Affairs. Closer inspection of this item revealed a possible mislabeling -- in any event the item seemed very person-oriented to the authors (this item involved the antonym of the word "peculiar").

The content of six of the nine verbal items biased towards males involve clear references to traditional or stereotyped male interests or skills: one involves the relationship between time and space,

transportation and communication; two are science items; one is a political science question; two are analogy items that deal with mechanical-electrical vocabulary, and one is a sentence completion dealing with war and man's labor and skills meeting human needs. This last item was perhaps most extreme in its disregard of females and at the same time the most statistically biased! Of the three verbal items biased in favor of women, two involve the antonyms of personality characteristics while the third involve a word found in cooking or as an adjective for clothing.

A check of the verbal discrete items as categorized by assembler's formal classifications (Table 6) revealed a pattern of average bias

---

Insert Table 6 about here

---

in which World of Practical Affairs and Science items are biased against women, whereas items coded as Aesthetic-philosophical and Human Relationships are biased, on the average, against males. These findings parallel Donlon's and Coffman's suggestions that items relating to "things" are easier for males, while items relating to "people" are easier for females. To make the comparison to Donlon and Coffman clearer, Table 6 also includes the mean difference between male and female deltas (item difficulty levels) by assembler's classification. As expected in view of the work of Donlon and Coffman, items relating to "things" (World of Practical Affairs and Science) are more difficult for females as compared to items relating to "people" (Aesthetic-Philosophical and Human Relationships), which

tend to be easier for females, although not to the same extent. In light of the average differences in difficulty level, the balance of items seems somewhat inappropriate. That is, 30 of the discrete verbale items are World of Practical Affairs or Science, whereas only 24 are coded as Aesthetic-Philosophical or Human-Relationships. An equalization of item types with the removal of the most biased items would be more equitable.

Analyzing the SAT-Verbal by item type (Table 7) demonstrated very slight bias, on the average, for any one item type. The sentence

---

Insert Table 7 about here

---

completion items were the most biased, on the average, and in a direction favoring males. Analogies also tended to favor males, while reading comprehension and antonyms apparently favor females very slightly. As the mean bias by item type is so slight, investigating items by content or classification code seems a more fertile approach for the verbal test. Donlon also found that reading comprehension and antonyms were easier for females, analogies for males, but found no difference for sentence completion.

Reference to male or female characters and pronoun usage did not seem to influence either the delta value or bias index (D-value) of the individual verbal items. Using the Tittle, et. al. ratio method, the verbal omnibus items have a male/female reference ratio of 1.75 for all usage (regular and generic) and a ratio of 1.11 for regular usage only, indicating an imbalance in favor of male references.

Moreover, reference was generally stereotypical. Throughout the items we find men in positions of power and fame -- in politics, music and art. No famous women were alluded to. Those female references made were to mothers and children, and of nymphs dancing alluringly for male gods -- positions of less status and power. Generic usage tended to appear limiting to female pursuits and could have been alleviated with words such as person, humanity, civilization, etc. More references were made to interests generally considered male than female -- for example, transportation, war, politics, mechanics, science, etc. Thus, the test seemed male-oriented... a finding that concurs with Donlon (1974). The twelve most biased items examined earlier also demonstrated this.

SAT-Mathematical Bias Analysis

Table 8 lists the item number, D-value and normalized D-value

---

Insert Table 8 about here

---

for all 60 math items. The delta plots for the math section appear in Figures 2-4. The math section was divided into its two component item types; regular math and data sufficiency. Although males outperform females on both math item types, this advantage is less marked for data sufficiency. The average delta difference (male minus female difficulty index) on the 18 data sufficiency items was -.6723 as contrasted to an average difference of -.8548 on the regular math items. Donlon (1974) also found an improved position for females on

the data sufficiency items. In terms of reducing sex bias, it would appear more equitable to have an equal number of data sufficiency and regular math items. Currently, there are 18 data sufficiency items and 42 regular math items. For a number of reasons, however, present plans call for eliminating the data sufficiency item type from future Scholastic Aptitude Tests.

The same criterion was used to determine item bias in the math sections as was used in the verbal test ($\pm$1.5 standard deviations distance). Items found biased by this criteria are starred in Table 8 and identified on the delta plots, Figures 2, 3, and 4. As can be seen by inspecting the starred items in Table 8, there is some variation as to which items meet the bias criterion, depending on whether the distributions for the entire math test or, for the component sections are used as the reference performance. As males and females tend to perform differently on data sufficiency items as compared to regular math items, it was decided that the separate distributions were fairer performance references.[5] By this method, a total of 5 of the 42 regular math items and 2 of the data sufficiency items were determined to be biased. These 7 items are listed in Table 9 with their D-values, normalized D-values, difficulty indexes, and assembler's code.

Insert Table 9 about here

---

[5]Our approach ignores item 35, which would have been considered in the combined analysis. For the curious, item 35 which seemed to be biased in favor of females when total math performance was considered as reference, is a data sufficiency algebra problem. However, due to its high difficulty level (Δ male = 17.5, Δ female = 17.9) and its position on the exam (last question in Section 7) it would probably have been discounted as a speed effect.

Of the five items biased in favor of males, three were coded as geometry and two as arithmetic. Because of the high difficulty level of item 32 (data sufficiency, arithmetic) and its near end position (Section IV has 35 items), it could be argued that this item is not really biased -- but merely answered rapidly and on the basis of partial information. Of the two items biased in favor of women, one was an algebra question and the other was coded " "Miscellaneous - elementary number theory." The miscellanceous item, item #2, involved "letter addition" -- filling in the missing units or tens digit in a two digit, four addend addition problem. Item 2 was of interest for several reasons: a) it is the only item in the entire math section which was easier for females than males, b) it proved so easy (94% of the females and 93% of the males answered it correctly) that it might be showing up as a biased item simply because it was so easy, and c) although coded as miscellaneous, it could have easily been solved by algebra (indeed, the authors did!).

The data was next subjected to an analysis for bias by content code. Table 10 gives the four assembler's content codes for the math section and the average bias for each, for both item types

---

Insert Table 10 about here

(regular math and data sufficiency) together, as well as separately.[6] Differences in average bias between regular math and data sufficiency items are to be expected, as the delta plots are calculated on the

[6] It should be noted that a total of five math items have been dropped from this analysis due to high delta values and their near end positions in their respective section. The eliminated items are: Section III -- 23 & 25 (both regular math); Section IV -- 32, 34, 35 (all data sufficiency)

basis of relative performance within a given section, and as females tend to have a slightly improved position on data sufficiency items. Examination of the total math section, however, suggests an interaction effect between item type and content code. That is, we find the average bias of a given content type also dependent on its item type. The geometry regular math items were the most biased, on the average, in favor of males. The data sufficiency geometry items were also biased in favor of males, but not to the same extent. Arithmetic regular math items were biased very slightly towards a male advantage, while arithmetic data sufficiency items showed a greater bias, but in the female direction. Algebra regular math items were biased, on the average, in favor of females, with algebra data sufficiency items showing almost no bias. The five miscellaneous regular math items demonstrated an average bias in favor of females, with the three miscellaneous data sufficiency items barely showing any bias, but still in a female direction. This is analogous to Donlon's analysis by content in which he found those items with algebra content to be easier, on the average, for females than geometry content items as compared to male performance on algebra and geometry items. It is interesting to speculate that the male advantage on geometry questions might have some relationship to the superior spatialization skills of males.

Donlon had designated a group of seventeen mathematics items as "subject content" -- items with real world referents, which he showed were more difficult for females as compared to males. Using the same

grouping technique, twelve items in the present study were found to have real world referents. Within these twelve items, eighteen male nouns or pronouns and eight females nouns or pronouns were used, generating a Tittle ratio of 2.25 (18/8). That is, references were made to males more than twice as often as to females. No generic pronouns were used. There did not seem to be a direct relationship between pronoun usage in a particular item and that item's bias direction. The references to the environment were somewhat mixed as to possibly stereotype, being not quite as male-oriented as in Donlon's study, six items involved male interests or skills, four involved female interests, and two were neutral. The gender characteristics of these mathematical items are very limited. For example, the test taker will find Maria's earnings exceeding last years, and two girls in a swimming pool; but must also calculate what percent of Jack's income is gross profit: the greatest number of items a boy can buy with his money, the temperature in an experiment: which of three men, Bill, Frank or Sam, have the most money: etc.

The item number, general content, item type, content type, delta value, and D-value are given in Table 11 for the twelve subject content items. For the seven regular math items with subject content,

---

Insert Table 11 about here

---

an average bias index of -.2749 was found. For the five data sufficiency items with subject content, an average bias index of -.0930 was generated. Thus, although data sufficiency items are generally

biased in favor of females, when only those data sufficiency items with subject content are investigated, even the data sufficiency items are biased, although only alightly, in favor of males. It is, however, difficult to define the existence or nonexistence of a relationship between specific stereotypical subject content and bias due to the small number of subject content items as compared to the number of variables needed to control for (e.g., item type, content type, pronoun usage). Thus, although the seven subject matter regular math items are the content group most biased against females, and the data sufficiency subject items were also biased against females, one cannot say it is due to their specific stereotypical content. No relationship could be assumed on the basis of the correlation coefficient between stereotypical male or female interest subject content and item bias assuming the data sufficiency and regular math groupings (content type, and pronoun usage were not controlled for).

It should be remembered that "bias" as defined here, is based on item performance relative to other items, not on absolute differences between the sexes. Items "biased" in favor of females may nonetheless be succeeded on more often by males.

Summary of Statistical Findings

Overall:

1) Performance on this form of the SAT-V was virtually equivalent for males and females.

2) On the average, males performed better than females on the SAT-M.

3) Within SAT-M, males performed better on both the regular math and data sufficiency items. Females do relatively better on the data sufficiency items as compared to their performance on regular math items.

Verbal Bias:

1) Twelve verbal items were found to be biased by the arbitrary criterion for values of D--nine (9) in favor of males, three (3) in favor of females. Six (6) of the nine (9) items biased in favor of males involved stereotypical male interests. Two of the three items biased in favor of females involved personality characteristics, a stereotypical female interest.

2) World of Practical Affairs and Science items are somewhat biased, on the average, in favor of males.

3) Aesthetic-Philosophical and Human-Relationship items are somewhat biased, on the average, in favor of females.

4) Analysis by verbal item type did not prove as fertile as did analysis by content. Sentence completion and analogy items were

slightly biased in the male favor. Reading comprehension and antonym items were slightly biased in the female direction.

5) More references were made to male nouns and pronouns than female. These references were generally stereotypical and offered a limited view of female pursuits.

6) Explicit reference in an item to male or female characters and the use of gender-oriented pronouns was not related to either the difficulty or bias index of the individual item.

Math Bias:

1) Five (5) regular math items and two (2) data sufficiency items were found to be biased--five in favor of males, two in favor of females.

2) Both regular math and data sufficiency geometry items were biased on the average in favor of males.

3) Algebra and miscellaneous regular math items were biased on the average in favor of females.

4) References were made to males more than twice as often as to females. There did not seem to be a direct relationship between the pronoun usage in a particular item and that item's bias.

5) Real world references were more mixed and not quite as male-oriented as in the verbal section.

6) Subject content real world reference items showed an average bias in favor of males.

7) A relationship between stereotypical content and statistical bias could not be shown for the 12 subject content items.

Educational significance

The content of test questions may determine which of two students of equal ability receives the higher score (Coffman, 1961; Donlon, 1973; Milton, 1958). Scores on aptitude tests, particularly the Scholastic Aptitude Test, partially determine both men's and women's access to higher education. Given the influence of aptitude tests and in the interests of social equality and fairness, it behooves us to investigate these tests for possible content bias.

It is hoped that this study will help make test users and test constructors more aware of possible sources of bias (e.g., item type, content, etc.) and aid in encouraging the development of unbiased, or at least balanced, tests. The findings should also be of value to the field practitioner--the test maker and the person who is guided by tests in the selection process.

The findings of virtual equivalence of male and female performance on the verbal section does not support the generalization of Maccoby and Jacklin (1974) in regard to female verbal superiority. The investigation of the relationship between item content and performance suggests that no single generalization can describe empirical outcomes when diverse groups are considered. In particular, no generalization which ignores test content is adequate, as content

may be related to performance. As Donlon (1973) suggests, "...other major educational tests should be examined to determine the conditions under which Maccoby's generalization is sustained" (p. 18).

References

Angoff, W. H. A technique for the investigation of cultural differences. Paper presented at the annual meeting of the American Psychological Association, Honolulu, September, 1972.

Angoff, W. H. & Ford, S. F. Item-race interaction on a test of scholastic aptitude. CEEB RDR-71-72, No. 3; RB-71-59. Princeton, N. J.: Educational Testing Service, 1971.

Angoff, W. H. & Stern, J. The equating of the scales for the Canadian and American Scholastic Aptitude Tests. CEEB RDR 7-72, No. 4; PR 71-24. Princeton, N. J.: Educational Testing Service, 1971.

Coffman, W. E. Sex differences in responses to items in an aptitude test. Eighteenth Yearbook, National Council on Measurement in Education, 1961, 117-124.

Donlon, T. F. Content Factors in sex differences on test questions. Research Memorandum 73-28. Princeton, N. J.: Educational Testing Service, 1973.

Donlon, T. F. & Angoff, W. H. The scholastic aptitude test. In W. H. Angoff (Ed.), The college board admissions testing program. Princeton, N. J.: College Entrance Examination Board & Educational Testing Service, 1971.

Lockheed-Katz, M. Sex bias in educational testing: a sociologist's perspective. Paper presented at the International Symposium on Educational Testing, The Hague, July, 1973.

Maccoby, E. E. & Jacklin, C. N. The psychology of sex differences. Stanford, Cal.,: Stanford University Press, 1974 (In press). Cited in Jacklin, C. N. & Maccoby, E. E. Mathematics, intellectual ability, and the sexes. Paper presented at the annual meeting of the American Educational Research Association, Chicago, April, 1974.

Milton, G. A. Five studies of the relation between sex-role identification and problem solving. Technical Report No. 3. New Haven, Conn.: Yale University, 1958.

Tittle, C. K., McCarthy, K. & Steckler, J. F. Women and educational testing. Princeton, N. J.: Educational Testing Service, 1974.

Walker, H. & Lev, J. Statistical inference. New York: Henry Holt & Co., 1954. Cited by W. E. Coffman, Sex differences in responses to items in an aptitude test. Eighteenth Yearbook, National Council on Measurement in Education, 1961, 117-124.

Table 1

Means and Standard Deviations of Raw (Formula) Scores and Item Deltas for the SAT-Verbal and Mathematical Sections

Sample	No. of cases	Test Score Data M	Test Score Data SD	Item Data Delta M	Item Data Delta SD
SAT-Verbal (90 Items)					
Male	1000	33.77	16.12	13.179	2.69
Female	1000	32.84	15.65	13.260	2.78
Total Population	499,307	33.3	16.0		
SAT-Mathematical (60 Items)					
Total Math (60 Items)					
Male	1000	26.39	13.01	12.56	2.47
Female	1000	21.83	12.12	13.36	2.66
Total Population	499,269	24.1	12.8		
Regular Math Items (42 Items)					
Male	1000	18.63	9.82	12.53	2.48
Female	1000	15.22	9.23	13.38	2.70
Data Sufficiency Items (18 Items)					
Male	1000	7.91	3.97	12.65	2.45
Female	1000	6.72	3.78	13.32	2.56

Table 2

Intercorrelations Between SAT Total Test Scores and Math Subscores for Male (N=1000) and Female (N=1000) Samples[a]

	Tot. Verbal	Tot. Math	Reg. Math	Data Suff.
Total Verbal		.6936	.6748	.5739
Total Math	.6787		.9729	.8625
Regular Math	.6606	.9788		.6787
Data Sufficiency	.5993	.8598	.7399	

[a] Correlations in upper triangle (◹) are for female sample.
Correlations in lower triangle (◺) are for male sample.

Table 3

Distribution of D-Values and Summary Data for the Item Delta Plots for the Male and Female Samples

D-Values	SAT-Verbal	SAT-Mathematical		
	Total Verbal	Total Math	Regular Math	Data Sufficiency
.80 - .899	-	-	-	-
.70 - .799	1	-	-	-
.60 - .699	2	-	-	-
.50 - .599	4	2	1	-
.40 - .499	5	1	1	-
.30 - .399	8	8	7	2
.20 - .299	11	7	4	2
.10 - .199	13	7	2	3
.00 - .099	11	6	6	4
-.10 - -.001	9	5	6	2
-.20 - -.101	5	8	5	0
-.30 - -.201	2	8	4	1
-.40 - -.301	1	3	3	2
-.50 - -.401	6	2	0	2
-.60 - -.501	4	1	1	-
-.70 - -.601	2	0	1	-
-.80 - -.701	1	1	1	-
-.90 - -.801	1	1	-	-
-1.00 - -.901	3	-	-	-
-1.099 - -1.000	1	-	-	-
-1.199 - -1.101	-	-	-	-
MIN	-1.024	-.815	-.759	-.450
MAX	.705	.530	.567	.342
R_{xy}	.979	.9868	.9864	.99
SD	.3959	.2938	.3006	.2478
N	90	60	42	18
M	0	0	0	0

Table 4

D-Values for SAT-Verbal Delta Plot (90 Items)

Section I Item #	D-Value	Normalized D-Value[a]	Item #	D-Value	Normalized D-Value[a]
1	.4676	1.1811	21	.5199	1.3131
2	.3516	.8881	22*	-.6294*	-1.5898*
3	-.1980	- .5000	23	.4992	1.2608
4*	-.9071*	-2.2910*	24*	.7053*	1.7814*
5	-.0307	- .0775	25	-.1918	- .4845
6	-.4450	-1.1240	26	-.0074	- .0187
7	.2231	.5636	27	.2138	.5399
8	.2241	.5661	28	.0366	.0925
9	-.5880	-1.4852	29	.2785	.7034
10	.1340	.3385	30	-.4746	-1.1988
11	.4339	1.0961	31	-.1467	- .3705
12	.2180	.5505	32	-.4704	-1.1882
13	.0335	.0847	33*	-.9666*	-2.4414*
14	-.4829	-1.2196	34	.2262	.5714
15	.0081	.0206	35	.1242	.3136
16	-.2928	- .7395	36	.1910	.4823
17	.5199	1.3131	37	-.5906	-1.4918
18	.1216	.3070	38	.0703	.1775
19	.0522	.1317	39*	-.6005*	-1.5167*
20	.1060	.2678	40	.1987	.5020

[a] The normalized D-Value is equal to $\frac{\text{D-Value}}{\text{SD of D-Values}}$. The standard deviation of D-Values for the Total Verbal section = .3959.

Table 4 (Continued)

Section II Item #	D-Value	Normalized D-Value[a]	Item #	D-Value	Normalized D-Value[a]
1	.3003	.7585	26	.1500	.3790
2	.0128	.0324	27	-.0881	- .2226
3	.2003	.5060	28	.2314	.5845
4	.0568	.1436	29	-.0053	- .0134
5	.1693	.4275	30*	-.8765*	-2.2138*
6	.3340	.8435	31	-.0022	- .0056
7	-.1441	- .3640	32*	-.9614*	-2.4284*
8	-.2829	- .7146	33	.1500	.3790
9	.0594	.1501	34	.0112	.0283
10	.1034	.2613	35	.1682	.4247
11*	-1.0236*	-2.5853*	36	.1335	.3372
12*	-.7382*	-1.8645*	37	.3676	.9285
13	-.0659	- .1665	38	.4111	1.0384
14	-.0530	- .1338	39	.3246	.8199
15	-.0457	- .1155	40	.5069	1.2804
16	.3034	.7663	41	-.4316	-1.0901
17	-.5518	-1.3937	42	.3526	.8905
18	.4458	1.1260	43	.0314	.0794
19	-.3093	- .7812	44	.2293	.5792
20*	.6189*	1.5631*	45	.2345	.5923
21	-.4347	-1.0979	46	.0822	.2077
22	-.1648	- .4163	47	.2392	.6041
23*	-.6445*	-1.6278*	48	-.0255	- .0644
24*	.6768*	1.7095*	49	.5017	1.2673
25	.3582	.9049	50	.1491	.3765

Table 5

D-Values, Delta Values, Assembler's Code, and Item Type for the 12 SAT-Verbal Outliers

	Item #	SAT Section	Orthogonal Distance: D-Value	Orthogonal Distance: Normalized D-Value	Delta: Male	Delta: Female	Assembler's Code	Item Type
Items biased in favor of males	4	I	-.9071	-2.2910	13.5	14.9	World of Prac. Aff.	Sentence Completion
	22	I	-.6294	-1.5898	13.5	14.5	World of Prac. Aff.	Antonym
	33	I	-.966	-2.4414	11.2	12.6	Biological Sci (Intended Inference)	R.C.
	39	I	-.6005	-1.5167	17.3	18.4	Synthesis (Application)	R.C.
	11	II	-1.0236	-2.5853	9.0	10.4	World of Prac. Aff.	Sentence Completion
	12	II	-.7382	-1.8645	9.3	10.3	Science	Sentence Completion
	23	II	-.6445	-1.6278	15.6	16.7	World of Prac. Aff.	Antonym
	30	II	-.8765	-2.2138	12.0	13.3	Science	Analogy
	32	II	-.9614	-2.4284	11.4	12.8	Science	Analogy
Items biased in favor of famales	20	II	.6189	1.5631	8.1	7.1	World of Prac. Aff.	Antonym
	24	I	.7053	1.7814	16.8	16.0	Aesthetic-Philo.	Antonym
	24	II	.6768	1.7095	15.7	14.9	Aesthetic-Philo.	Antonym

Table 6

Analysis of SAT-Discrete Verbal Items

by Assembler's Formal Classifications

Classification	Frequency	Mean D-Value	Mean Normalized D-Value	Mean Δ Difference (F-M)
World of Practical Affairs	15	-.2471	-.6241	-.4533
Science	15	-.1370	-.3461	-.2733
Aesthetic-Philosophical	13	.2017	.5095	.1846
Human Relationships	11	.1230	.3106	.1000

Table 7

Analysis of SAT-Verbal by Item Type

Item Type	Frequency	Mean D-Value	Mean Normalized D-Value
Sentence Completion	18	-.0855	-.2159
Antonym	18	.0390	.0985
Analogy	19	-.0587	-.1482
Reading Comprehension	35	.0558	.1408
	Total = 90		

Table 8

D-Values for Delta Plots of SAT-Mathematical

Total Math (N=60)					
Item #	D-Value	Normalized D-Value[a]	Item #	D-Value	Normalized D-Value[a]
Section III			Section IV		
1	-.2801	- .9533	1*	-.5521*	-1.8788*
2*	.4638*	1.5784*	2	-.1070	- .3640
3	-.2590	- .8814	3	.2015	.6857
4	-.3586	-1.2204	4	-.0027	- .0091
5	-.2016	- .6862	5	-.4269	-1.4529
6	-.1443	- .4910	6	.0336	.1142
7	.1175	.3999	7	.0495	.1683
8	-.1026	- .3490	8	.3373	1.1480
9	-.0134	- .0455	9	.3003	1.0219
10	.0489	.1665	10	-.0559	- .1902
11	.3107	1.0574	11	-.0976	- .3322
12	.3318	1.1293	12	.1697	.5774
13	-.1078	- .3668	13	-.2859	- .9728
14	.1120	.3813	14	.3368	1.1462
15*	-.8149*	-2.7734*	15	.3368	1.1462
16*	-.7524*	-2.5604*	16	-.1346	- .4582
17	.1955	.6652	17	-.0144	- .0492
18	-.4069	-1.3846	18	.3324	1.1312
19*	.5039*	1.7149*	19	-.3849	-1.3100
20	.2056	.6998	20	-.3010	-1.0243
21	-.1240	- .4218	21	.1540	.5242
22	-.2708	- .9214	22	.1329	.4523
23	.2262	.7698	23	.2690	.9155
24	-.1975	- .6721	24	.0390	.1328
25	-.2137	- .7272	25	.2901	.9874
			26	.0808	.2748
			27	.0959	.3262
			28	.2742	.9332
			29	.1806	.6147
			30	-.1758	- .5984
			31	.3733	1.2704
			32	-.2710	- .9224
			33	-.2181	- .7422
			34	.2419	.8231
			35*	.5297*	1.8027*

[a] Standard deviation of all Math items = .2938.

[b] Standard deviation of Reg. Math items = .3006.

[c] Standard deviation of Data Suff. items = .2478.

Table 8 (Continued)

Regular Math (N=42)			Data Sufficiency (N=18)		
Item #	D-Value	Normalized D-Value[b]	Item #	D-Value	Normalized D-Value[c]
Section III			Section IV		
1	-.2633	- .8759	18	.3054	1.2322
2*	.4523*	1.5048*	19*	-.4358*	-1.7585*
3	-.2501	- .8323	20	-.3293	-1.3286
4	-.3430	-1.1412	21	.0859	.3468
5	-.1835	- .6107	22	.0453	.1829
6	-.1169	- .3891	23	.1836	.7406
7	.1414	.4705	24	-.0117	- .0471
8	-.0685	- .2279	25	.2242	.9045
9	.0122	.0407	26	.0137	.0552
10	.0960	.3194	27	-.0327	- .1320
11	.3544	1.1790	28	.1867	.7534
12	.3675	1.2226	29	.1577	.6361
13	-.0746	- .2481	30	-.2533	-1.0219
14	.1465	.4873	31	.2469	.9962
15*	-.7588*	-2.5247*	32*	-.4500*	-1.8156*
16*	-.6862*	-2.2830*	33	-.3345	-1.3496
17	.2433	.8096	34	.0560	.2258
18	-.3420	-1.1379	35	.3419	1.3796
19*	.5673*	1.8874*			
20	.2665	.8868			
21	-.0705	- .2347			
22	-.2068	- .6880			
23	.3018	1.0043			
24	-.1331	- .4429			
25	-.1291	- .4295			
Section IV					
1*	-.5448*	-1.8127*			
2	-.1078	- .3588			
3	.2161	.7190			
4	.0133	.0441			
5	-.3995	-1.3293			
6	.0667	.2220			
7	.0738	.2456			
8	.3624	1.2059			
9	.3422	1.1387			
10	-.0029	- .0097			
11	-.0514	- .1709			
12	.2020	.6720			
13	-.2471	- .8223			
14	.3846	1.2797			
15	.3846	1.2797			
16	-.0715	- .2380			
17	.0566	.1884			

Table 9

D-Values, Delta-Values, and Assembler's Content Code

for the Seven SAT-Mathematical Outliers by Item Type

Item Type	Item #	Section	Orthogonal Distance		Delta		Assembler's Content Code
			D-Value	Normalized D-Value[a]	Male	Female	
Regular Math							
Biased in favor of males	15	III	-.7588	-2.5247	13.9	16.0	Geometry (Circles, rotation of polygons)
	16	III	-.6862	-2.2830	15.1	17.2	Arithmetic (Percent)
	1	IV	-.5448	-1.8127	8.5	9.8	Geometry (Polygons, not inscribed or circumscribed)
Biased in favor of females	2	III	.4523	1.5048	7.1	6.8	Miscellaneous (Ele. number theory, letter addition)
	19	III	.5673	1.8874	15.7	16.0	Algebra (Systems of equations and inequalities)
Data Sufficiency							
Biased in favor of males	19	IV	-.4358	-1.7585	10.4	11.6	Geometry (Circles, rotation of polygons)
	32	IV	-.4500	-1.8156	16.5	18.0	Arithmetic (Properties of integers, number judgment)

[a] Standard deviation of Regular Math items = .3006
Standard deviation of Data Sufficiency items = .2478

Table 10

Mean Bias by Assembler's Classification Code and Item Type

for SAT-Mathematical Items

Classification Code	Item Type	Frequency	Items Eliminated[a]	Mean D-Value
Geometry				
	All	18		-.1214
	Reg. Math	11	Sect. III-23 & 25	-.1772
	Data Suff.	4	Sect. IV-34	-.1094
	Total	15	Sect. III-23 & 25; Sect. IV-34	-.1627
Arithmetic				
	All	17		-.0221
	Reg. Math	12	None	-.0260
	Data Suff.	4	Sect. IV-32	.1111
	Total	16	Sect. IV-32	-.0066
Algebra				
	All	17		.0889
	Reg. Math	12	None	.0970
	Data Suff.	4	Sect. IV-35	-.0108
	Total	16	Sect. IV-35	.0614
Miscellaneous				
	All	8		.1311
	Reg. Math	5	None	.1850
	Data Suff.	3	None	.0293
	Total	8	None	.1311

[a] Items were eliminated on the basis of high delta-values and near end position for a given section.

Table 11

Content, Item Type, Assembler's Content Code, Delta-Values, and D-Values for the 12 Subject-Content Math Items

Item #	Subject Content	Item Type	Content Code	Delta Male	Delta Female	D-Value
Section III						
4	Mary, hours in school	Reg. Math	Arith.	9.6	10.7	-.3430
5	No. of pencils in X boxes, each containing Y pencils	Reg. Math	Algebra	10.0	10.5	-.1835
16	Jack reselling flags, % profit	Reg. Math	Arith.	15.1	17.2	-.6862
24	Article with greatest value per cubic inch	Reg. Math	Arith.	15.3	16.6	-.1331
Section IV						
2	How many set price items can boy buy with his money	Reg. Math	Algebra	7.9	8.5	-.1078
5	Time and temperature during science experiment	Reg. Math	Algebra	10.9	12.2	-.3995
16	% of students taking French & Art	Reg. Math	Misc.	15.2	16.4	-.0715
Section IV						
20	Capacity in cups of pitcher	Data Suff.	Algebra	9.4	10.4	-.3295
24	Rate of interest for one year	Data Suff.	Arith.	10.7	11.3	-.0117
27	Maria's earnings exceeding last years	Data Suff.	Algebra	14.4	15.2	-.0327
29	Which of three men--Bill, Sam, or Frank--has the most money	Data Suff.	Misc.	9.5	9.8	.1577
30	Faster girl swimmer passing slower girl swimmer	Data Suff.	Misc.	11.8	12.8	-.2533

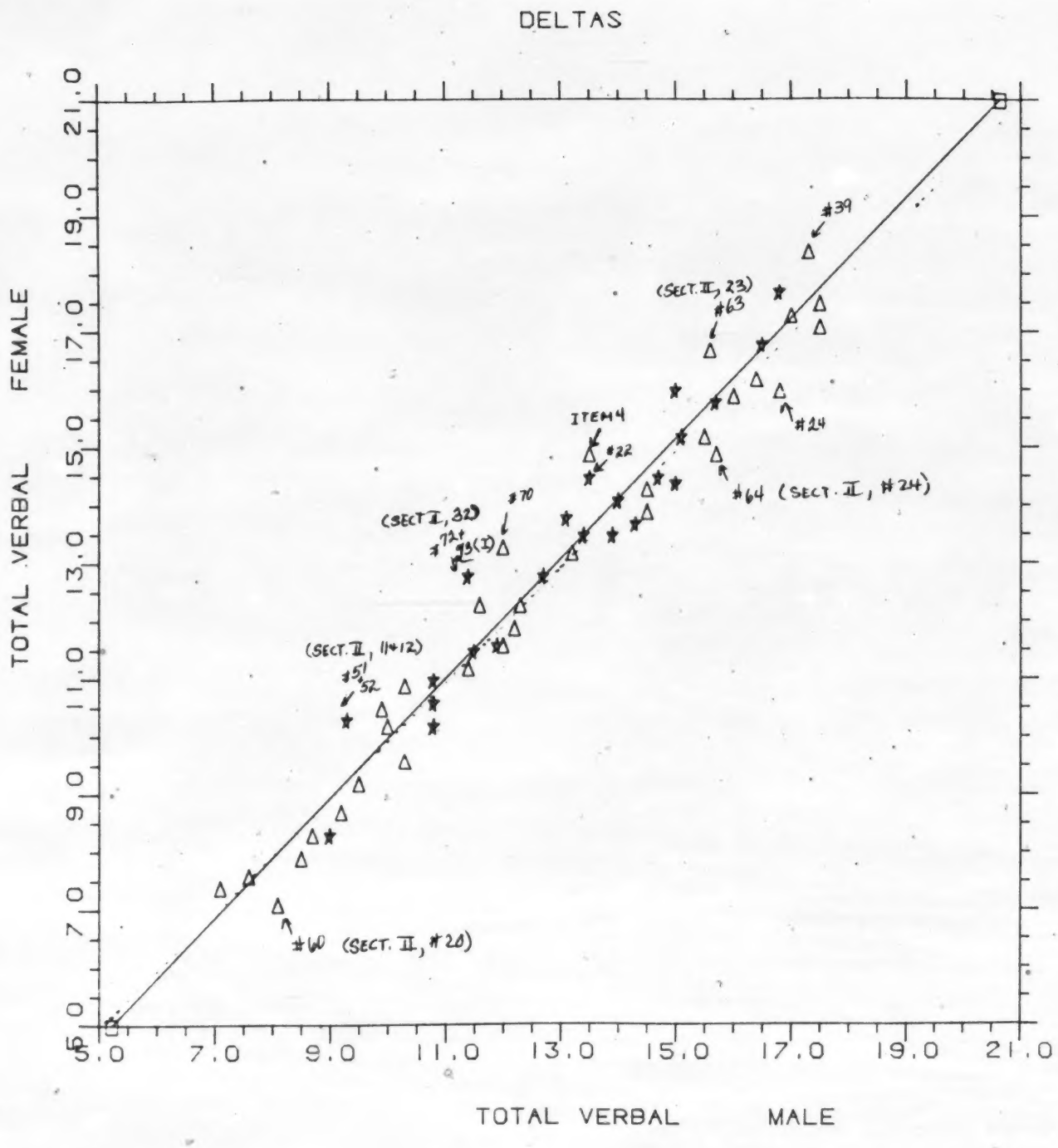

Figure 1. Delta Plot of SAT Total Verbal Section (n=90 items) for Male and Female Sample.

r = .979

△ △ △ SINGLE POINTS
★ ★ ★ MULTIPLE POINTS

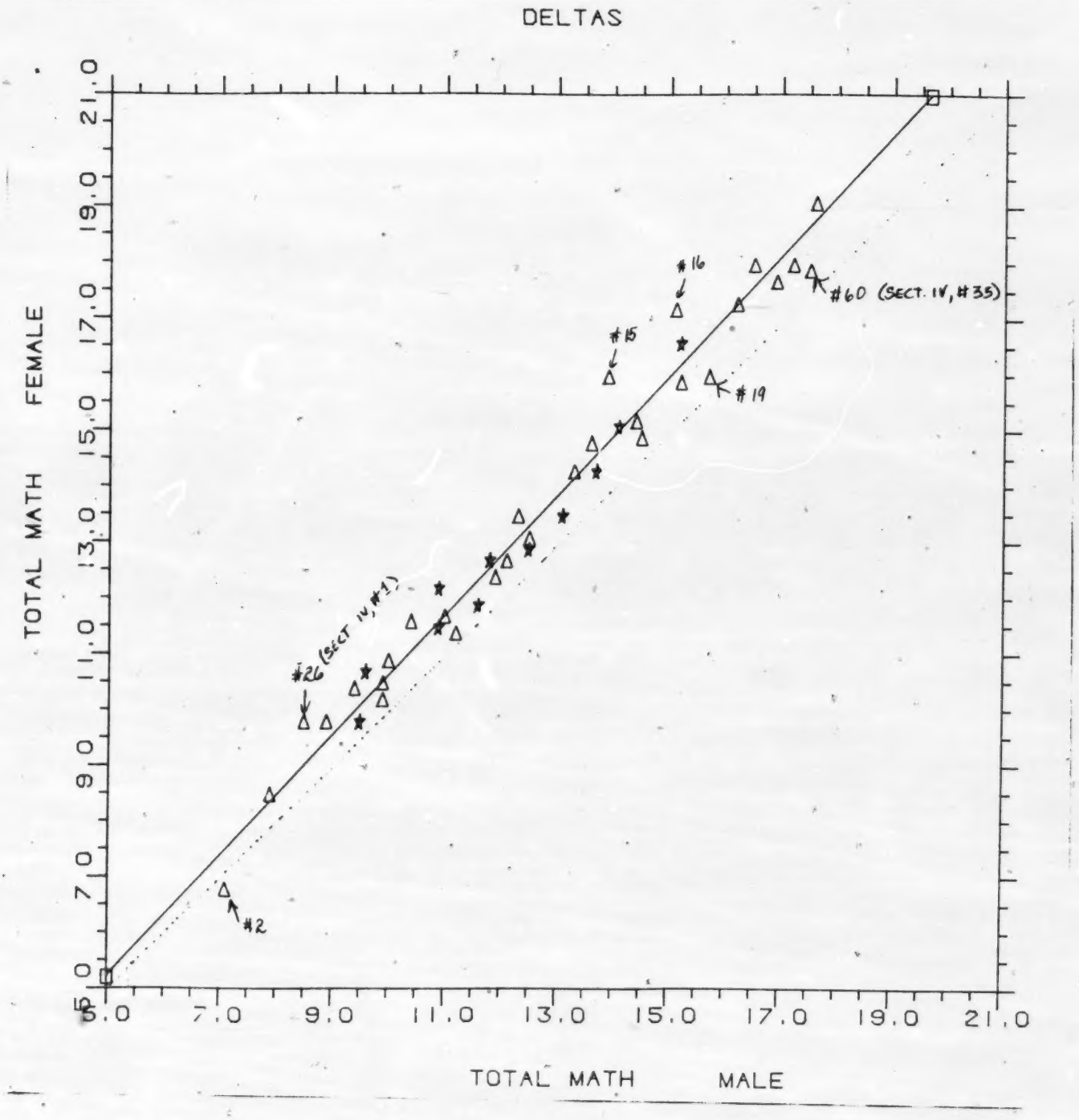

Figure 2. Delta Plot of SAT Total Math Section (n = 60 items) for Male and Female Sampl

r = .9868

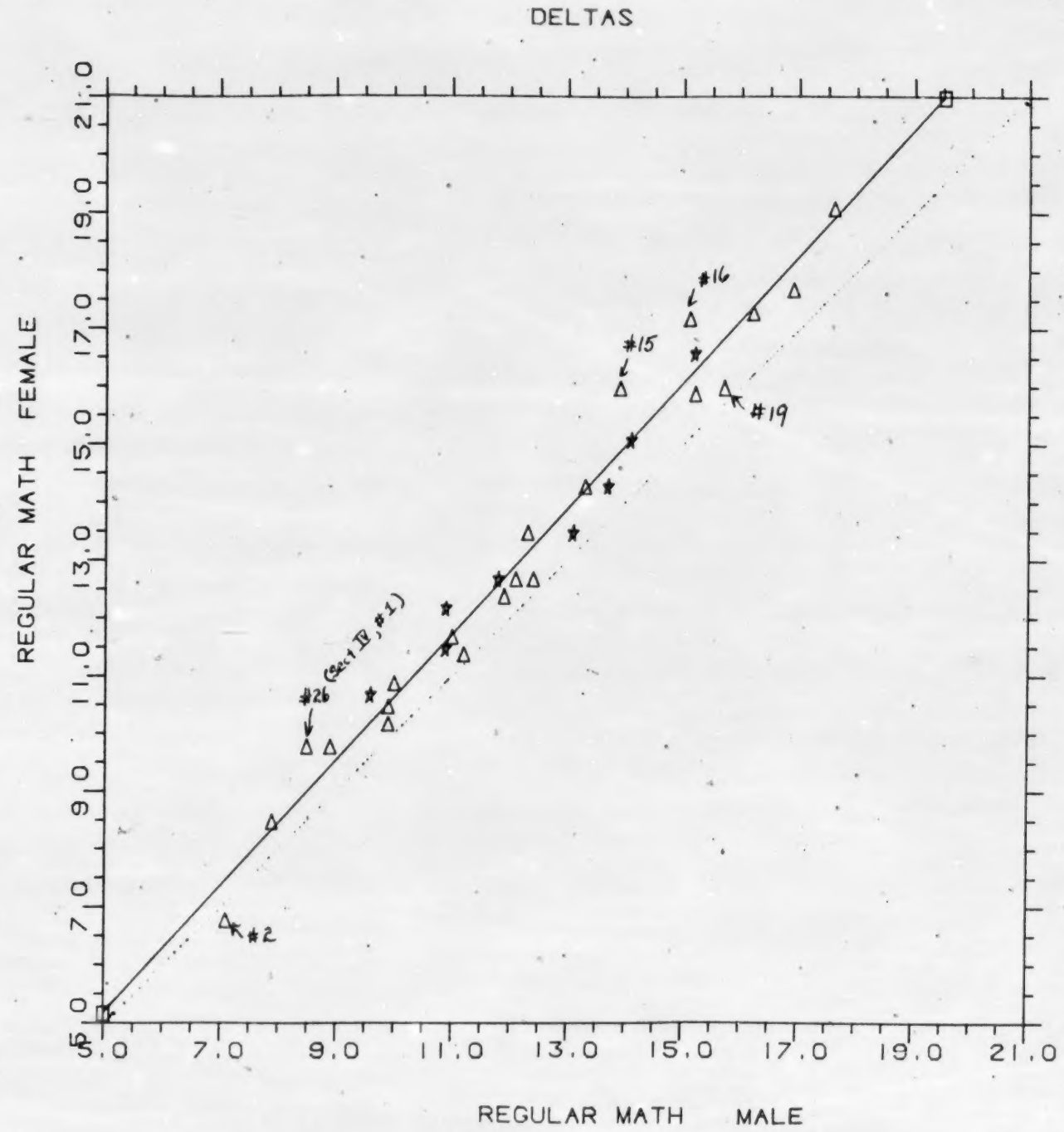

Figure 3. Delta Plot of SAT Regular Math Items (n=42) for Male and Female Sample.

r=.9864

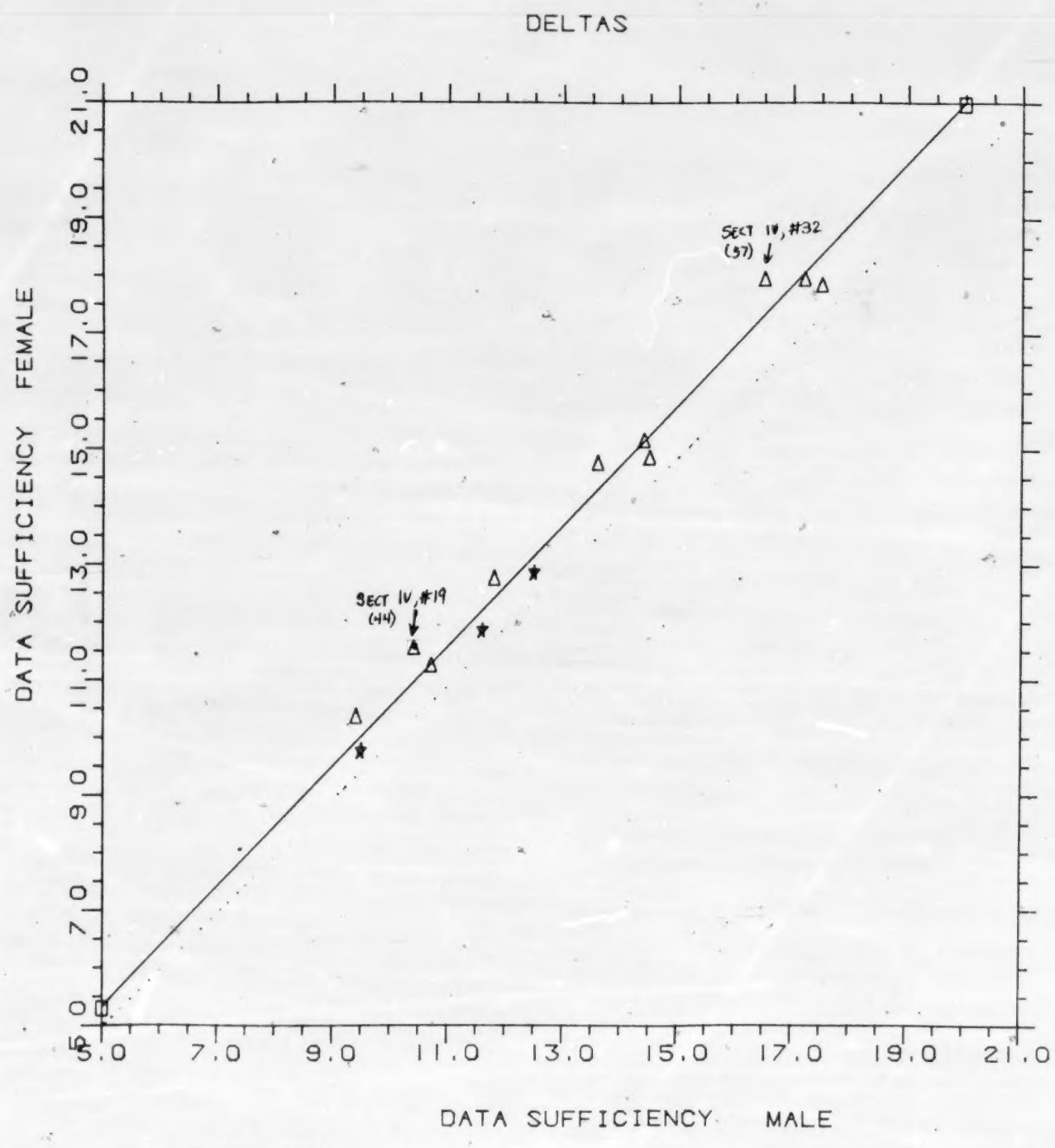

Figure 4. Delta Plot of SAT Data Sufficiency Items (n=18 items) for Male and Female Sample.

r = .99